

PATENT  
81940.0071

Express Mail Label No. EV 324 112 150 US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

Masahiro ARAI et al.

Serial No: Not assigned

Filed: February 20, 2004

For: Magnetic Disk Array Device With  
Processing Offload Function Module

Art Unit: Not assigned

Examiner: Not assigned

**TRANSMITTAL OF PRIORITY DOCUMENT**

Mail Stop PATENT APPLICATION

Commissioner for Patents

P.O. Box 1450

Alexandria, VA 22313-1450

Dear Sir:

Enclosed herewith is a certified copy of Japanese patent application No. 2003-393912 which was filed November 25, 2003, from which priority is claimed under 35 U.S.C. § 119 and Rule 55.

Acknowledgment of the priority document(s) is respectfully requested to ensure that the subject information appears on the printed patent.

Respectfully submitted,

HOGAN & HARTSON L.L.P.

Date: February 20, 2004

By: 

Anthony J. Orler

Registration No. 41,232

Attorney for Applicant(s)

500 South Grand Avenue, Suite 1900  
Los Angeles, California 90071  
Telephone: 213-337-6700  
Facsimile: 213-337-6701



日 本 国 特 許 庁  
JAPAN PATENT OFFICE

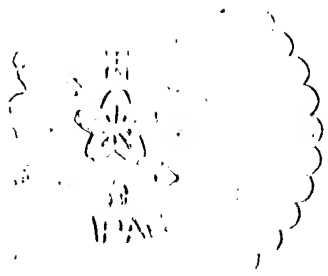
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日                      2 0 0 3 年 1 1 月 2 5 日  
Date of Application:

出 願 番 号                      特 願 2 0 0 3 - 3 9 3 9 1 2  
Application Number:  
[ST. 10/C]:                      [ J P 2 0 0 3 - 3 9 3 9 1 2 ]

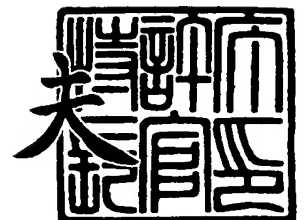
出 願 人                      株式会社日立製作所  
Applicant(s):



2 0 0 3 年 1 2 月 2 5 日

特許庁長官  
Commissioner,  
Japan Patent Office

今 井 康



出証番号    出証特 2 0 0 3 - 3 1 0 7 4 3 8

【書類名】 特許願  
【整理番号】 KN1577  
【提出日】 平成15年11月25日  
【あて先】 特許庁長官殿  
【国際特許分類】 G06F 12/00  
【発明者】  
    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所  
                                システム開発研究所内  
    【氏名】 新井 政弘  
【発明者】  
    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所  
                                システム開発研究所内  
    【氏名】 松並 直人  
【発明者】  
    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所  
                                システム開発研究所内  
    【氏名】 八木沢 育哉  
【発明者】  
    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所  
                                システム開発研究所内  
    【氏名】 萬年 暁弘  
【発明者】  
    【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 R  
                                A I D システム事業部内  
    【氏名】 佐藤 雅彦  
【特許出願人】  
    【識別番号】 000005108  
    【氏名又は名称】 株式会社 日立製作所  
【代理人】  
    【識別番号】 100093492  
    【弁理士】  
    【氏名又は名称】 鈴木 市郎  
    【電話番号】 03-3591-8550  
【選任した代理人】  
    【識別番号】 100078134  
    【弁理士】  
    【氏名又は名称】 武 顕次郎  
    【電話番号】 03-3591-8550  
【手数料の表示】  
    【予納台帳番号】 113584  
    【納付金額】 21,000円  
【提出物件の目録】  
    【物件名】 特許請求の範囲 1  
    【物件名】 明細書 1  
    【物件名】 図面 1  
    【物件名】 要約書 1

**【書類名】 特許請求の範囲****【請求項 1】**

A T A 磁気ディスクと、前記 A T A 磁気ディスクを制御するディスクアレイコントローラと、前記ディスクアレイコントローラと前記 A T A 磁気ディスクとのパス上に存在するインターフェースカード内の処理オフロード機能モジュールと、を備えた磁気ディスクアレイ装置であって、

前記ディスクコントローラは、リード又はライトを一例とする標準処理を行う標準処理 F C コマンドと、ベンダユニークなオフロード処理を行うオフロード処理 F C コマンドと、を前記インターフェースカードに出力し、

前記インターフェースカード内の前記処理オフロード機能モジュールは、前記標準処理 F C コマンドに対しコマンドマッピングテーブルを用いて対応する A T A コマンドを前記 A T A 磁気ディスクに発行し、前記オフロード処理 F C コマンドに対し A T A プロトコルで最適処理となる A T A コマンド群を用意し必要なときには演算する

ことを特徴とする磁気ディスクアレイ装置。

**【請求項 2】**

請求項 1 に記載の磁気ディスクアレイ装置において、

前記処理オフロード機能モジュールは、コマンド解析処理部と演算処理部を備え、

前記コマンド解析処理部において、前記ディスクアレイコントローラからの標準処理 F C コマンドが A T A コマンドにマッピング可能であるか否かを判断し、

前記演算処理部において、前記オフロード処理 F C コマンドを前記コマンド解析部から引き継いで前記 A T A コマンド群を用意する

ことを特徴とする磁気ディスクアレイ装置。

**【請求項 3】**

請求項 1 又は 2 に記載の磁気ディスクアレイ装置において、

前記ディスクコントローラは、前記磁気ディスクアレイ装置の内部に前記処理オフロード機能モジュールの存在有無と利用可能なオフロード処理の判別を行うオフロード処理判別部を備える

ことを特徴とする磁気ディスクアレイ装置。

**【請求項 4】**

請求項 2 に記載の磁気ディスクアレイ装置において、

前記演算処理部は、前記オフロード処理 F C コマンドがパリティ演算付きライトであるときに、A T A 磁気ディスクから読み出すリードコマンド発行と、パリティ演算の実行と、A T A 磁気ディスクへ書き込むライトコマンド発行と、を行う

ことを特徴とする磁気ディスクアレイ装置。

**【請求項 5】**

請求項 2 に記載の磁気ディスクアレイ装置において、

前記演算処理部は、前記オフロード処理 F C コマンドがオンライン複数ディスクベリファイであるときに、前記コマンド解析処理部から前記オフロード処理 F C コマンドを引き継いで、リストとして渡された I D に対応する磁気ディスクにリードコマンドを同時発行する

ことを特徴とする磁気ディスクアレイ装置。

**【請求項 6】**

A T A 磁気ディスクと、前記 A T A 磁気ディスクを制御するディスクアレイコントローラと、の間のパス上に存在するインターフェースカード内の処理オフロード機能モジュールであって、

前記処理オフロード機能モジュールは、リード又はライトを一例とする標準処理を行う標準処理 F C コマンドと、ベンダユニークなオフロード処理を行うオフロード処理 F C コマンドと、を前記ディスクコントローラから受け取り、

前記標準処理 F C コマンドに対しコマンドマッピングテーブルを用いて対応する A T A コマンドを前記 A T A 磁気ディスクに発行し、前記オフロード処理 F C コマンドに対し A

T A プロトコルで最適処理となる A T A コマンド群を用意し必要なときには演算することを特徴とする処理オフロード機能モジュール。

【請求項 7】

請求項 6 に記載の処理オフロード機能モジュールにおいて、  
前記処理オフロード機能モジュールは、コマンド解析処理部と演算処理部を備え、  
前記コマンド解析処理部において前記ディスクアレイコントローラからの標準処理 F C コマンドが A T A コマンドにマッピング可能であるか否かを判断し、  
前記演算処理部において前記オフロード処理 F C コマンドを前記コマンド解析部から引き継いで前記 A T A コマンド群を用意することを特徴とする処理オフロード機能モジュール。

【請求項 8】

請求項 7 に記載の処理オフロード機能モジュールにおいて、  
前記オフロード処理 F C コマンドは、パリティ演算付きライトコマンド、オンライン複数ディスクベリファイコマンド、R A I D フォーマットコマンド、又はディスク間コピーコマンドである  
ことを特徴とする処理オフロード機能モジュール。

【請求項 9】

A T A 磁気ディスクと、前記 A T A 磁気ディスクを制御するディスクアレイコントローラと、前記ディスクアレイコントローラと前記 A T A 磁気ディスクとのパス上に存在するインターフェースカードと、を備えた磁気ディスクアレイ装置であって、  
前記 A T A 磁気ディスク毎に処理オフロード機能モジュールを接続したディスク収納容器を 1 つ以上設け、  
前記ディスクコントローラは、リード又はライトを一例とする標準処理を行う標準処理 F C コマンドと、ベンダユニークなオフロード処理を行うオフロード処理 F C コマンドと、を前記インターフェースカードに出力し、  
前記処理オフロード機能モジュールは、前記標準処理 F C コマンドに対しコマンドマッピングテーブルを用いて対応する A T A コマンドを前記 A T A 磁気ディスクに発行し、前記オフロード処理 F C コマンドに対し A T A プロトコルで最適処理となる A T A コマンド群を用意することを特徴とする磁気ディスクアレイ装置。

【請求項 10】

請求項 9 に記載の磁気ディスクアレイ装置において、  
前記処理オフロード機能モジュールは、コマンド解析処理部と演算処理部を備え、  
前記コマンド解析処理部において、前記ディスクアレイコントローラからの標準処理 F C コマンドが A T A コマンドにマッピング可能であるか否かを判断し、  
前記演算処理部において、前記オフロード処理 F C コマンドを前記コマンド解析部から引き継いで前記 A T A コマンド群を用意することを特徴とする磁気ディスクアレイ装置。

【請求項 11】

請求項 9 又は 10 に記載の磁気ディスクアレイ装置において、  
前記オフロード処理 F C コマンドは、パリティ演算付きライトコマンド、オンライン複数ディスクベリファイコマンド、R A I D フォーマットコマンド、又はディスク間コピーコマンドである  
ことを特徴とする磁気ディスクアレイ装置。

【請求項 12】

磁気ディスクと、前記磁気ディスクを制御するディスクアレイコントローラと、前記ディスクアレイコントローラと前記磁気ディスクとのパス上に存在するインターフェースカード内の処理オフロード機能モジュールと、を備えた磁気ディスクアレイ装置であって、  
前記磁気ディスクは、A T A 磁気ディスクと F C 磁気ディスクとが混在配置され、  
前記ディスクコントローラは、リード又はライトを一例とする標準処理を行う標準処理

FC コマンドと、ベンダユニークなオフロード処理を行うオフロード処理FC コマンドと、を前記インターフェースカードに出力し、

前記インターフェースカード内の前記処理オフロード機能モジュールは、前記ディスクコントローラからのFC コマンドが、前記標準処理FC コマンドであり且つFC 磁気ディスクを対象としたコマンドであれば、当該FC コマンドに対して処理を加えずに素通りさせて該当するFC 磁気ディスクに送り、前記標準処理FC コマンドであり且つATA 磁気ディスクを対象としたコマンドであれば、当該FC コマンドに対してコマンドマッピングテーブルを用いて対応するATA コマンドを前記ATA 磁気ディスクに発行し、

前記処理オフロード機能モジュールは、前記ディスクコントローラからのFC コマンドが、前記オフロード処理FC コマンドであれば、対象とする磁気ディスクの種別と処理内容に応じた最適なコマンド群に変換し処理を行う

ことを特徴とする磁気ディスクアレイ装置。

【請求項 13】

請求項 12 に記載の磁気ディスクアレイ装置において、

前記処理オフロード機能モジュールは、コマンド解析処理部と演算処理部を備え、

前記コマンド解析処理部において、前記ディスクアレイコントローラからのFC コマンドが、標準処理FC コマンドであり且つATA 磁気ディスクを対象としたコマンドであれば、ATA コマンドにマッピング可能であるか否かを判断し、

前記演算処理部において、前記ディスクコントローラからのFC コマンドが前記オフロード処理FC コマンドであれば、前記オフロード処理FC コマンドを前記コマンド解析部から引き継いで最適なコマンド群に変換し処理を行う

ことを特徴とする磁気ディスクアレイ装置。

【請求項 14】

磁気ディスクと、前記磁気ディスクを制御するディスクアレイコントローラと、前記ディスクアレイコントローラと前記磁気ディスクとのパス上に存在するインターフェースカードと、を備えた磁気ディスクアレイ装置であって、

前記磁気ディスクは、ATA 磁気ディスクとFC 磁気ディスクとが混在配置され、

前記ディスクアレイコントローラはコントローラ処理部を有し、前記コントローラ処理部に接続された処理オフロード機能モジュールが前記ディスクアレイコントローラ毎に設置され、

前記コントローラ処理部は、リード又はライトを一例とする標準処理を行う標準処理FC コマンドと、ベンダユニークなオフロード処理を行うオフロード処理FC コマンドと、を前記処理オフロード機能モジュールに出力し、

前記ディスクアレイコントローラ内の前記処理オフロード機能モジュールは、前記コントローラ処理部からのFC コマンドが、前記標準処理FC コマンドであり且つFC 磁気ディスクを対象としたコマンドであれば、当該FC コマンドに対して処理を加えずに素通りさせて該当するFC 磁気ディスクに前記インターフェースカードを介して送り、前記標準処理FC コマンドであり且つATA 磁気ディスクを対象としたコマンドであれば、当該FC コマンドに対してコマンドマッピングテーブルを用いて対応するATA コマンドを前記ATA 磁気ディスクに前記インターフェースカードを介して発行し、

前記処理オフロード機能モジュールは、前記コントローラ処理部からのFC コマンドが、前記オフロード処理FC コマンドであれば、対象とする磁気ディスクの種別と処理内容に応じた最適なコマンド群に変換し処理を行う

ことを特徴とする磁気ディスクアレイ装置。

**【書類名】 明細書****【発明の名称】** 処理オフロード機能モジュールを備えた磁気ディスクアレイ装置**【技術分野】****【0 0 0 1】**

本発明は、主としてコンピュータの外部記憶装置システムに関わり、特に、磁気ディスクアレイ装置の性能向上を図る技術に関する。

**【背景技術】****【0 0 0 2】**

従来技術として、複数のディスクにデータの処理を分散させ且つ冗長データを記録することによって、高信頼で高速にデータを記録する磁気ディスクアレイ装置（RAID装置：Redundant Arrays of Inexpensive Disks）が広く用いられている。磁気ディスクアレイの詳細については、例えば非特許文献1に記載されている。

**【0 0 0 3】**

一般に、磁気ディスクアレイ装置は、多数のディスクを搭載しており、これらの磁気ディスクに適切にデータを分散格納し処理するために、入出力制御装置であるディスクアレイコントローラが搭載されている。

**【0 0 0 4】**

ディスクアレイコントローラは、データを適切なディスクに格納するための処理を行う機能の他に、複数のディスクをRAID向けにフォーマットする機能や、ディスクの万一の故障においてもデータを失うことのないように冗長データを計算して書き込む機能や、また、故障予兆のあるディスクを未然に検出する予防保守機能などを有していて、磁気ディスクへの入出力の他にも多数の制御処理を行っている。このため、ディスクアレイコントローラには高速にデータを処理する性能が求められており、改良が進められてきた。

**【0 0 0 5】**

一方で、磁気ディスクアレイ装置の高性能化を図る別の手法として、全ての処理をディスクアレイコントローラに一任するのではなく、配下に処理の一部を肩代わりさせるサブコントローラを設け、これらのサブコントローラに処理を任せるオフロード処理（ディスクコントローラからロード（負荷）をオフにする処理）という方法がある。

**【0 0 0 6】**

オフロード処理により処理高速化を図る方法としては、例えば、特許文献1において、ファイバチャネルループを複数のループに分断し、ループごとに置かれるサブコントローラを用いることで、複数のフォーマットを多重実行する方法が開示されている。

**【0 0 0 7】**

この特許文献1には、データの流れるファイバチャネルループを分断することによって、ループ上を流れるデータを局所化し、更に、コントローラのデータバスの使用効率を改善することが可能である旨が開示されている。また、コントローラが行う処理の一部をサブコントローラに行わせることによって、コントローラのI/O処理負荷を軽減することが可能である旨が開示されている。さらに、ディスク間のコピーや、装置を停止させずにオンライン中にディスクの表面チェックを行う、予防保守の一種であるオンラインベリファイなどの処理も可能である旨が開示されている。

**【特許文献1】** 特開 2 0 0 1 - 2 2 5 2 6 号公報**【非特許文献1】** 「A Case for Redundant Arrays of Inexpensive Disks (RAID)」, David A. Patterson, Garth Gibson, and Randy H. Katz, CoMPuTer Science Division Department of Electrical Engineering and CoMPuTer Sciences, University of California Berkeley**【発明の開示】****【発明が解決しようとする課題】****【0 0 0 8】**

磁気ディスクアレイ装置に対する要望としては、磁気ディスクアレイ装置の高性能化が求められる一方で、データの爆発的増加を背景に業務用途やデータの種類・重要度に応じ

て、格納先となる磁気ディスクに従来のような高いサーバ用ディスクではなくて、より安価なディスクに保存したいという利用者のニーズが生まれてきている。このようなニーズに基づいて、近年では主に個人向けパソコン（PC）などで多用されてきたATA（ATA Attachment）規格磁気ディスク（以下、ATAディスクと称する）を、企業向け磁気ディスクアレイ装置でも利用する動きがある。

#### 【0009】

ATAディスクは、ホストのインターフェースコントローラ部に当たる機能をディスク内に内蔵し、従来から主にサーバ用途で多用されてきたSCSIディスクやFC（Fibre Channel）ディスクに比べて、複雑な回路を必要とせず、安価に製造が可能なディスクである。

#### 【0010】

また、ATA規格の論理仕様である制御コマンドは、ディスクコントローラが複雑な判断をせずに済むように、機能の微妙な違いごとに異なるコマンドが割り当てられている。例えば、ライト系コマンドでは、一般のWRITEコマンドのほか、DMA（Direct Memory Access）が設定されている場合にはDMA転送でライト行うWRITE DMA、セクタ単位での書き込みを行うWRITE SECTORSなどがあり、さらに、それぞれのコマンドに、WRITE（RETRY）、WRITE（NORETRY）のように失敗時にリトライ処理を行うか否かで2種類の別々のコマンドが用意される。

#### 【0011】

一方、物理的な仕様においては、元来多数のディスクを接続する必要がないPC内蔵用として設計されているため、パラレルケーブルを用いて接続するParallel ATAディスクでは1バスあたりに最大2台しか接続できず、インターフェース高速化のためParallel ATAに代わって登場してきたSerial ATA（以下、SATAと称する）では、Point to Point接続形態により1バスに1台しか接続できない。また、コントローラとディスク間のケーブルも1m以下と制限がある。このようなことから、多数の磁気ディスクを複数の磁気ディスク格納筐体にわたって格納し、さらに、これらの磁気ディスクをディスクアレイコントローラと接続する必要がある磁気ディスクアレイ装置では、接続形態ならびに接続距離においてATAディスクをそのまま適用するのは難しい。そのため、磁気ディスクアレイ装置では従来より使われてきたファイバチャネル（以下、FCと称する）にコマンドや物理インターフェースを変換して接続するFC-ATA変換接続方式が用いられている。

#### 【0012】

このFC-ATA変換接続方式は、ATAコマンドを対応するFCコマンドにマッピングすることでFC-ATA論理変換を実現し、さらに、ATAとFCの信号レベルなどの物理インターフェースを変換することで、ディスクアレイコントローラからはATAディスクをFCディスクであるかのように扱えるようにするものである。

#### 【0013】

FC-ATA論理変換部においては、2つのプロトコル間で類似するコマンドをマッピングすることとなるが、2つの規格はそれぞれ別の団体が策定した規格であり、まったく異なる仕様を持つ。このため、備える機能もおのずと違い、実際にはFCディスクの備える機能に対して、ATAディスクには該当する機能がないものもある。このような機能呼び出すFCコマンドに対しては、ATAコマンドをマッピングすることはできない。

#### 【0014】

従って、ディスクアレイコントローラでは、コマンド発行に先立ち、まずターゲットがFC-ATA変換によって、FCコマンドを受け付けられるATAディスクなのか、単なるFCディスクなのかを判別する必要がある。その後、ターゲットがATAディスクであった場合、発行しようとするコマンドがFC-ATA変換で変換可能なコマンドであるかを判別し、さらに、変換不可能なコマンドであった場合には、他のFCコマンドの組み合わせで、ほぼ同等の処理を行えるようにするか、もしくは、コマンドの発行を取りや



めるかを選択する必要がある。その結果、ディスクアレイコントローラでは判別のための処理が必要となり、処理に要する負荷が増大し、結果として性能を落としてしまう可能性がある。

#### 【0015】

一方、マッピングされるATAコマンドから見た場合、FCのコマンドにマッピングできるのは、FCコマンドに仕様としてあるものだけである。上述したように、ATAは目的の違いに応じて同種のコマンドが複数存在する。しかし、FC-ATA変換においてディスクアレイコントローラから発行されるFCコマンドを択一的にマッピングするしかなく、豊富なATAのコマンドを十分に活かしきれない。また、FC規格上に該当する機能がないものは、そもそもマッピングすることができないという課題が生じる。

#### 【0016】

上述したような背景をもつATAディスクを用いた磁気ディスクアレイ装置に対して、従来技術のようなオフロード処理の適用を考えた場合に次のような課題が生じる。

#### 【0017】

第1の課題は、コマンドオーバーヘッドの増加である。従来技術のようにサブコントローラを設けた場合、オフロード処理によってディスクアレイコントローラの処理は分散できるものの、ディスクアレイコントローラからの指示は、サブコントローラでFCコマンドでの処理に展開され、次に、FC-ATA変換部でATAコマンドに変換され、そして、ATAディスクへと伝播するという冗長な経路を取る。即ち、ディスクアレイコントローラ、サブコントローラ、FC-ATA変換部という構成要素が縦続的に接続される経路を構成している。また、この処理に対する応答も同じ経路をたどるため、結果としてコマンドに対するディスクの応答が悪化してしまうおそれがある。

#### 【0018】

第2の課題は、FC-ATA変換では必ずしも最適なATAコマンドが選択されないという課題を依然として解決できない点である。前述の特許文献1によれば、サブコントローラはFCループ上にあるため、発行できるコマンドもおのずとFCコマンドとなってしまう。このため、FC-ATA変換における最適なATAコマンドの選択に課題が残る。

#### 【0019】

第3の課題は、IDリソース消費による接続可能なディスク台数の減少である。サブコントローラがディスクコントローラから指示を直接受けるためにはFCループ上のデバイスとして認識できる必要があり、この認識のためにループIDを消費する。サブコントローラを多数設置すればそれだけ多重化が可能となるが、この多数設置の分だけIDが消費されるため、接続できるディスク台数は減少し、大容量化のニーズに相反することになる。

#### 【0020】

そこで、本発明の目的は、ATAディスクの利用を可能としながら、FC-ATAコマンド変換によるオーバーヘッドを削減し、最適なATAコマンドを活用し、さらに、ID等のリソースを無消費にして、オフロード処理を実現する磁気ディスクアレイ装置を提供することにある。

#### 【課題を解決するための手段】

#### 【0021】

前記課題を解決するために、本発明は主として次のような構成を採用する。

ATA磁気ディスクと、前記ATA磁気ディスクを制御するディスクアレイコントローラと、前記ディスクアレイコントローラと前記ATA磁気ディスクとのパス上に存在するインターフェースカード内の処理オフロード機能モジュールと、を備えた磁気ディスクアレイ装置であって、

前記ディスクコントローラは、リード又はライトを一例とする標準処理を行う標準処理FCコマンドと、ベンダユニークなオフロード処理を行うオフロード処理FCコマンドと、を前記インターフェースカードに出力し、

前記インターフェースカード内の前記処理オフロード機能モジュールは、前記標準処理

FCコマンドに対しコマンドマッピングテーブルを用いて対応するATAコマンドを前記ATA磁気ディスクに発行し、前記オフロード処理FCコマンドに対しATAプロトコルで最適処理となるATAコマンド群を用意し必要なときには演算する構成とする。

【発明の効果】

【0022】

本発明によれば、ATAディスクの利用を可能としながらも、コマンド変換によるオーバーヘッドの増加を最小限にし、かつ、ディスクの論理インターフェースに合わせた最適なコマンド変換を実現し、ID等のリソースを無消費にしてオフロード処理を可能とすることができる。

【0023】

さらに、論理インターフェースの異なる複数種の磁気ディスクを同一の上位コマンドで操作することが可能である。

【0024】

また、全てのオフロード処理用コマンドは、対象となる磁気ディスクに対して通常（標準）コマンドと同じように発行する形式をとることから、たとえば受け取ったデータを暗号化又は復号化してリード又はライトを行うオフロード処理機能を用意し、すべてのリード又はライトをこの機能と呼び出すベンダユニークコマンドによって行えば、容易に暗号化機能付きディスクアレイ装置を提供することも可能である。

【発明を実施するための最良の形態】

【0025】

「第1の実施形態」

図1は、本発明の第1の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールを磁気ディスク格納筐体のインターフェースカードに実装した構成を示す図である。図面において、以下、xは任意の整数を表す。

【0026】

磁気ディスクアレイ装置101（以下、ディスクアレイと称する）は、SANクライアント120xからSAN（Storage Area Network）111と呼ばれるネットワークを通じてアクセスされる二次記憶装置である。SANクライアント120xはSAN111を経由せず直接接続することも可能である。磁気ディスクアレイ装置101は、ATA磁気ディスク17xxに対しデータ転送制御を行うディスクアレイコントローラ102xと、複数のATA磁気ディスクを格納する磁気ディスク格納筐体104xと、を備えている。ディスクアレイ格納筐体104xとディスクアレイコントローラ102xは、ファイバチャネル（以下、FCと称する）113x、114xで接続される。

【0027】

磁気ディスク格納筐体104xは、複数のATA磁気ディスク17xxとインターフェースカード105xとから構成されており、ディスクアレイコントローラ102xは前記インターフェースカード105xを通じて磁気ディスク17xxにデータ転送を行う。インターフェースカード17xxは、ディスクアレイコントローラ102xと後続する磁気ディスク格納筐体とを相互に接続するFC接続部115（例えば、スイッチなど）と、図示するようにFC接続部115に接続される処理オフロード機能モジュール106と、から構成される。

【0028】

尚、図1では、ディスクアレイコントローラは磁気ディスク格納筐体の外部に配置されているが、本発明においては、ディスクアレイコントローラが磁気ディスク格納筐体内にあっても構わない。また、ディスクアレイコントローラにおいては図示の例では2つ記載されているが、1つでも又は3つ以上存在しても構わない。

【0029】

次に、図2はディスクアレイコントローラ102xの内部構成を示す図である。フロントエンド接続インターフェースコントローラ201はFCケーブル108xを通じ、SA

N111と接続され、データ転送やコマンドデータの送受信を行う。バックエンド接続インターフェースコントローラ207xはATA磁気ディスク間とデータ送受信を行うためのものである。205は中央処理制御部であり、メモリ206に格納されたディスクアレイ制御プログラム2061を動作させる。メモリ206は前記ディスクアレイ制御プログラム2061の他に、ディスク情報取得プログラム2062とディスク情報管理テーブル2063とを備える。ディスク情報取得プログラム2062は起動時に接続されたATA磁気ディスク17xxの情報を取得するためのものであり、内部に処理オフロード機能モジュール（以下、機能モジュールと称する）106の有無と利用可能なオフロード処理の判別を行うオフロード処理判別部103xを備える。ディスクアレイ制御プログラム2061はオフロード処理判別部103xに加えて、オフロード処理用のベンダユニークコマンド発行部2064を有する。

#### 【0030】

202はデータ転送コントローラであり、ディスクアレイ制御プログラム2061やディスク情報取得プログラム2062の指示により、フロントエンド接続インターフェースコントローラ201、バックエンド接続インターフェースコントローラ207x、データバッファ203間のデータ移動を行う。

#### 【0031】

管理ネットワークコントローラ204は、ディスクアレイコントローラ102xの管理用LAN109xを通じ、管理用端末110から各種監視やディスク構成や設定の変更を行うために用いられる。管理用端末は管理用LAN109xを経由せず直接接続されていてもよい。

#### 【0032】

次に、図3は本発明の第1の実施形態に関する処理オフロード機能モジュールの内部構成を示す図である。302xはFC接続インターフェースであり、203xを介してFC接続部115と接続される。FCインターフェースコントローラ303は、FC接続インターフェース302xを通じてデータバッファ301を用いてデータの送受信を行う。同様に、308xはATA接続インターフェースであり、ATA接続ライン310xを用いてATA磁気ディスクドライブ17xxと接続され、ATAインターフェースコントローラ306はデータバッファ301とATA接続インターフェース308x間のデータ送受信を行う。

#### 【0033】

MPU305は、メモリ307に格納されたプログラム等を用いて演算処理やコマンド変換、インターフェースコントローラ303、306の制御を行う処理演算装置である。3091はFCのIDとそれに対応するATA磁気ディスクの関連付けを管理するためのディスクID情報管理テーブルである。ATA磁気ディスクはIDという概念はなく、直接にFC IDを持つことができない。そこで、このテーブル3091において、FCのIDとATA接続インターフェースNoとの対応関係を記録することにより、間接的にFC IDの割り当てを実現する。また同様に、ATAディスクはFCディスクが持つディスク固有の識別子であるWWNをもてないため、WWNとATA磁気ディスクの固有番号である製造番号との対応付けについてもこのテーブル3091で管理する。

#### 【0034】

3092はオフロード処理時の各コマンドの実行状況を管理するコマンド実行状況管理テーブルであり、コマンドとそのコマンドの実行ステータス、ならびにデータの受信があった場合にはそのバッファのアドレスが記録される。3093はコマンド解析処理プログラム、3094はステータス処理プログラム、3095はATAコマンドマッピングテーブルであり、3096は演算処理プログラムである。

#### 【0035】

次に、図4は図2に示すディスク情報管理テーブル2063の詳細構造を示す図である。401は磁気ディスクアレイ装置101で管理される磁気ディスクの番号、402は磁気ディスクの製造社名（ベンダ名）であり、403は製品名、404は各磁気ディスクに

固有の製造番号である。4 0 5 はその磁気ディスクの容量を表す。また、4 0 7 は機能モジュール 1 0 6 によって提供される処理オフロード機能であり、各磁気ディスクに対してどの機能が利用可能であるかを表す。図示するオフロード機能としては、パリティ演算付きライト、R A I D フォーマット、複数ディスクベリファイ、ディスクコピーの機能がある。なお、情報の取得の仕方については後述する。このディスク情報管理テーブルについては、図示するよりも、より多くの項目があってもよいし、項目の順序や名称、記載方法に関して異なってもよい。

#### 【0 0 3 6】

次に、本発明の第 1 の実施形態に係る磁気ディスクアレイ装置におけるディスク情報取得処理の動作について、図 5 を参照しながら以下説明する。図 5 は、ディスクアレイコントローラ 1 0 2 x のディスク情報取得プログラム 2 0 6 2 が、当該コントローラ 1 0 2 x で使用する A T A 磁気ディスク 1 7 x x を認識し、図 4 に示したディスク情報管理テーブル 2 0 6 3 に情報登録を行う動作を示したフローチャートである。破線枠 5 0 1 で囲まれた処理はディスクアレイコントローラ 1 0 2 x での動作を示し、同様に、5 0 2 は処理オフロード機能モジュール（機能モジュール）1 0 6 での動作を示し、5 0 3 は磁気ディスクでの動作を示す。

#### 【0 0 3 7】

ディスクアレイコントローラ 1 0 2 x は磁気ディスクアレイ装置 1 0 1 の電源が投入され、コントローラ内部の初期設定が終了すると、ディスク情報取得プログラム 2 0 6 2 を起動する（5 0 0 0）。

#### 【0 0 3 8】

ディスク情報取得プログラム 2 0 6 2 は、接続されている全ディスクの状態を取得するため、ディスクの構成情報をチェックする F C コマンドである I n q u i r y コマンドと、磁気ディスクを一意に特定可能な F C I D を指定して、データバッファ 2 0 3 に用意する。なお、ここでは、F C I D は 0 から始まるものとする（5 0 0 1、5 0 0 2、5 0 0 3）。データバッファ 2 0 3 にデータが用意できると、当該プログラム 2 0 6 2 はデータ転送コントローラ 2 0 2（図 2 を参照）とバックエンド接続インターフェースコントローラ 2 0 7 x に指示を出し（5 0 0 3）、1 1 3 x ないし 1 1 4 x を通じて、コマンドを発行する（5 0 0 4）。

#### 【0 0 3 9】

発行されたコマンドは、各磁気ディスク格納筐体 1 0 4 x のインターフェースカード 1 0 5 x を介して伝達する。ディスクアレイコントローラ 1 0 3 x は該当 I D をもつディスクに向かってコマンドを発行しているが、実際には A T A ディスクが F C I D を直接持つことはできない。そのため、機能モジュール 1 0 6 がその I D とディスクの対応付けを管理しており、該当 I D を持つ磁気ディスクを管理する機能モジュール 1 0 6 がディスクに代わってコマンドを受領する（5 0 0 6）

機能モジュール 1 0 6 がコマンドを受領すると、機能モジュール 1 0 6 内にある F C インターフェースコントローラ 3 0 3 は、モジュール内データバッファ 3 0 1 にそのコマンドを格納し（5 0 0 7）、割り込みを用いて M P U 3 0 5 に通知する（5 0 0 8）。M P U 3 0 5 に通知が来ると、コマンド解析処理プログラム 3 0 9 3 はモジュール内データバッファ 3 0 1 からコマンドを読み込み、解析を行う（5 0 0 9）。解析により当該プログラム 3 0 9 3 はコマンドが A T A のコマンドにマッピング可能かどうか判断する（5 0 1 0）。

#### 【0 0 4 0】

マッピング可能な場合、コマンド解析処理プログラム 3 0 9 3 は、コマンドマッピングテーブル 3 0 9 5 を用いて対応する A T A コマンドを選択し、この選択したコマンドと、ディスク I D 情報管理テーブル 3 0 9 1 から F C I D と対応によって得られる A T A 接続インターフェース N o（A T A 接続インターフェース N o と A T A ディスク N o とは対応付けされている）と、必要であればコマンドの実行により得られるデータを格納するモジュール内データバッファ 3 0 1 の格納先アドレスの 3 つを用意し、前記データバッファ

301内に用意する(5011)。もし、マッピング可能でなかった場合には、コマンド解析処理プログラム3093は演算処理プログラム3096に処理を移し、適切な処理を行う(5012)。なお、演算処理プログラム3096に処理を移した場合の詳細な動作については別の処理において説明する。

#### 【0041】

コマンドが前記データバッファ301に用意できると、コマンド解析処理プログラム3093はATAインターフェースコントローラ306に指示をして、モジュール内データバッファ301で指定されたNoをもつATA接続インターフェースを介してATA磁気ディスクにコマンドを発行する(5014)。

#### 【0042】

コマンドに対しディスクが応答すると(5016)、応答はATA接続インターフェース308xを介して、ATAインターフェースコントローラ306によってデータバッファ301に格納され(5018)、格納したことをMPU305に通知する(5019)。

#### 【0043】

以下、上述した手順とは逆に、ステータス処理プログラム3094は該当する応答内容から、FCコマンド用のステータスを作成し(5020)、バッファに格納(5021)し、FCインターフェースコントローラ303に指示を出し(5022)、ステータスをディスクアレイコントローラ103xに報告する(5023)。以上の一連の動作により、ディスクアレイコントローラ103xはコマンドの結果を得るが、ディスクアレイコントローラ103xからは直接磁気ディスクが応答したかのように見えるため、処理過程において、処理オフロード機能モジュールの仲介はディスクアレイコントローラ103xにとって透過的である。

#### 【0044】

報告を受けた磁気ディスクアレイ装置は、同様の手順によって、READ CAPACITYコマンドでディスクの容量を取得する(5025)。また、ベンダユニークなコマンドによって、該当ディスクに対し利用可能なオフロード処理機能についての情報を問い合わせ(5026)、得られた情報をディスク情報管理テーブル2063に登録していく(5027)。なお、ベンダユニークコマンドの処理に関しては後述する。1台のディスクが終わると、最後のIDまで同様の処理を繰り返す(5028, 5029)。

#### 【0045】

以上のようにしてディスク情報を登録するが、全てのディスクに対して情報を取得できるのであれば、他の方法でも良い。また、1コマンドで機能モジュールに接続される全ディスクの情報を取得するようなベンダユニークコマンドを用意し、そのベンダユニークコマンドによって情報取得してもかまわない。また、図4のディスク情報管理テーブル2063に示した管理情報は一例であり、項目内容や情報や登録表記方法が異なっても構わない。

#### 【0046】

次に、本発明の第1の実施形態に係る磁気ディスクアレイ装置において、複数のディスクに異なるコマンドを送るオフロード処理について以下説明する。複数のディスクに異なるコマンドを送るオフロード処理例として、パリティ演算付きライトの動作を説明する。図6は、図5に示す処理5010において演算処理プログラムが呼び出された場合の処理5012を示したものである。

#### 【0047】

ディスクアレイコントローラ102xは、ホスト120xからコマンドとライトするデータを受け取ると、ディスクアレイ制御プログラム2061における処理の過程上でパリティ演算が必要かどうかを判断する。演算が必要であった場合、ディスクアレイ制御プログラム2061はディスク情報管理テーブル2063を参照し、データを書き出す磁気ディスク17xxが該当オフロード処理を実行可能かどうか判別する。判別の結果、実行可能な場合は、当該処理を指示するベンダユニークなコマンドを生成し、図5に示すステッ

プ 5 0 0 3 ~ 5 0 0 5 の動作によりライトを行うディスクのうちの 1 つに対して発行する。コマンドは、コマンド識別番号、データを書き出すディスクの ID とパリティを書き出すディスクの ID からなる ID リスト、ライトするデータの長さ、などの情報を持つ。なお、リストにおいてデータ書き出すディスクの ID は複数あってよい。

#### 【 0 0 4 8 】

コマンドを発行されたディスクを管理する機能モジュール 1 0 6 がコマンドを受領すると、前述した図 5 のステップ 5 0 0 7 の動作によりコマンドとライトする新データをモジュール内データバッファ 3 0 1 に格納する。その後、ステップ 5 0 0 9 の動作により、コマンド解析処理プログラム 3 0 9 3 はコマンド解析を開始する。このコマンド 3 0 9 3 はオフロード処理を行うベンダユニークコマンドであるため、図 5 に示すステップ 5 0 1 0 、 5 0 1 2 の動作により、その処理を演算処理プログラム 3 0 9 6 に引き継ぐ。

#### 【 0 0 4 9 】

演算処理プログラム 3 0 9 6 の目的は、与えられたオフロード処理を達成するために A T A プロトコルで最適な処理となるようコマンド群を用意し、必要であれば演算を行い、目的の処理を達成し、結果をディスクアレイコントローラ 1 0 2 x に報告することである。このコマンド 3 0 9 6 で行う処理は、新データと旧データ並びに旧パリティデータから演算によって新パリティデータを算出し、該当ディスクへ新データ、新パリティデータを書き出すことである。この処理を大きく分けると、旧データ及び旧パリティの読み出し（処理 A）、新パリティの算出（処理 B）、新データ及び新パリティの書き込み（処理 C）の大きく 3 段階に分かれる。

#### 【 0 0 5 0 】

演算処理プログラム 3 0 9 6 は、まず、旧データ及び旧パリティの読み出し処理（処理 A）を実現するために、コマンド解析処理プログラム 3 0 9 3 が解析した結果から、データを書き出すディスクの ID 群とパリティを書き出すディスクの ID を取得し、ディスク ID 情報管理テーブル 3 0 9 1 より、対応する A T A 接続インターフェース No を順に取得する（6 0 0 1）。次に、処理に最適な A T A リードコマンドを選択し（6 0 0 2）、このコマンドと、前記 A T A 接続インターフェース No と、リードしたデータを格納するためのモジュール内バッファ 3 0 1 のデータ格納アドレスとを、データないしパリティのリードが必要なディスク台数分繰り返して、コマンド実行状況管理テーブル 3 0 9 2 に登録する（6 0 0 3、6 0 0 4）。ここで、処理に最適な A T A リードコマンドには、前述した旧データ及び旧パリティの読み出し処理が旧データと旧パリティを高速かつ確実に読み上げることであることから、高速転送可能でかつリードに失敗すると再試行を試みられる READ DMA (RETRY) コマンドを用いる。コマンドの登録が終了したらこの管理テーブル 3 0 9 2 の情報を、モジュール内データバッファ 3 0 1 に転送し（6 0 0 5）、コマンド実行状況管理テーブル 3 0 9 2 の各コマンドの処理状況を「コマンド発行待ち」に設定する（6 0 0 6）。

#### 【 0 0 5 1 】

モジュール内データバッファ 3 0 1 にコマンドの登録が終わると演算処理プログラム 3 0 9 6 は、A T A インターフェースコントローラ 3 0 6 に指示をして前述した一連のコマンドを発行させる（6 0 0 7）。A T A インターフェースコントローラ 3 0 6 はコマンドの発行が終わると、割り込みによって発行処理終了を知らせる（6 0 0 8）。演算処理プログラム 3 0 9 6 は発行終了割り込みを受けると、コマンド実行状況管理テーブル 3 0 9 2 に登録したコマンドの処理状況を、「応答待ち」に変更する（6 0 0 9）。

#### 【 0 0 5 2 】

その後、当該磁気ディスク 1 7 x x がコマンドを受領し、データの転送を終了し、応答ステータスを返すと（6 0 1 1、6 0 1 2）、A T A インターフェースコントローラ 3 0 6 は応答ステータスを受領したことを、割り込みによって演算処理プログラム 3 0 9 6 に通知する（6 0 1 3）。演算処理プログラム 3 0 9 6 はモジュール内データバッファ 3 0 1 からステータスを読み取り、成功していればコマンド実行状況管理テーブル 3 0 9 2 における該当コマンドの処理状況を「完了」にし、リードしたデータが格納されているデー

タバッファアドレスをコマンド実行状況管理テーブル 3092 に登録し、失敗していればエラーとして、そのエラー内容を処理状況に登録する (6014)。

#### 【0053】

コマンド実行状況管理テーブル 3092 上のコマンドの処理状況がすべて「完了」になった場合、演算処理プログラムは、前述した3段階の内の新パリティの算出処理に移行する。もし、このテーブル 3092 上に1つでもエラーがある場合には、演算処理プログラムはエラーレポートを作成し、ディスクアレイコントローラ 102x にエラー通知を行い、以降、新パリティの算出処理、新データ及び新パリティの書き込み処理に移らずに、オフロード処理を終了する (6017)。エラーレポートは、エラー通知とともにディスクアレイコントローラ 102x に通知される。なお、コマンド実行状況管理テーブル 3092 上のコマンドに「応答待ち」が残っていた場合、一定時間を待って、さらに応答が帰ってこない場合には当該コマンドはタイムアウトとしてエラー処理する (6015, 6016, 6017)。

#### 【0054】

3段階の内の新パリティの算出処理 (処理B) に移行すると、演算処理プログラム 3096 は、あらかじめ指定したデータバッファのアドレスに格納されていた新データと旧データ及び旧パリティの読み出し処理によって取得した旧データ、旧パリティデータより排他的演算処理を行い、新パリティデータを算出する (6019)。なお、パリティデータは他の方法で算出されてもかまわず、また、図6ではパリティデータは1つとしているが、複数のパリティデータがあってもよい。

#### 【0055】

次に、3段階の内の新データ及び新パリティの書き込み処理 (処理C) に移行し、最適なATAライトコマンドを選択し、ライトを行う全ディスクに対してコマンドと該当ディスクが接続されている接続インターフェースNo、書き込みデータが格納されているバッファのアドレスをコマンド実行状況管理テーブル 3092 に登録し (6020, 6021, 6022)、次に、旧データ及び旧パリティの読み出し処理と同様に、モジュール内データバッファ 301 にコマンドを格納してATAインターフェースコントローラ 306 にコマンド発行を指示し、ディスクからの応答を待つ (6100)。パリティ演算でのライトは確実に高速かつ書き込めることが望まれるため、ATAコマンドにはWRITE DMA (RETRY) を利用する。

#### 【0056】

演算処理プログラム 3096 はATAインターフェースコントローラ 306 を通じて新データと新パリティを無事書き込めたという報告を受け取ると、FCステータスを作成し (6030)、ディスクアレイコントローラ 102x に報告を行い (6033)、オフロード処理を終了する。

#### 【0057】

次に、本発明の第1の実施形態に係る磁気ディスクアレイ装置において、複数のディスクに対して同一のコマンドを送る処理について以下説明する。複数のディスクに対して同一のコマンドを送る処理例として、オンラインベリファイ処理の動作を説明する。ここで、オンラインベリファイとは、ホストからのI/Oとは非同期でディスクの表面チェックを行い、故障予兆のあるディスクを検出することを目的とする保守作業である。

#### 【0058】

通常、ディスクアレイコントローラでオンラインベリファイを行う際は、各ディスクに対しVERIFYコマンドをディスク全体がチェックし終わるまで複数回発行し、さらにこれをディスク台数分繰り返す必要がある。本発明では、オフロード処理用コマンドとして、コマンド識別番号と同時実行したいディスクのIDをリストにしたものを、代表ディスクにのみ1コマンドを発行するだけで、対象とする全てのディスクに対して、同時にオンラインベリファイ処理を実行可能するものである。以下にその動作を説明する。

#### 【0059】

ディスクアレイコントローラ 102x は、動作経過時間などの或る契機によってオンラ

インベリファイが必要となったディスクがあると、実行対象となるディスクのIDをリストにする。ここで対象となるディスクは単数でもよいが、ここでの説明では複数を想定する。ディスクアレイコントローラ102xはオンラインベリファイ処理の必要な磁気ディスクのFC IDのリストとコマンド識別番号を用いて、オンラインベリファイオフロード処理実行用のベンダユニークなコマンドを作成し、代表ディスクに発行する。ここで、代表ディスクとはリストに渡されるIDを割り当てられたディスクのうちの1つである。なお、ディスクアレイコントローラがコマンドを発行する手順については、図5のステップ5003～5005の動作と同じく、データバッファ203にコマンドと送信先IDを格納し、データ転送コントローラ202と、バックエンド接続インターフェースコントローラ207xに指示することで行う。

#### 【0060】

当該IDを管理する機能モジュール106がコマンドを受け取ると、コマンド解析処理プログラム3093は受け取ったコマンドがオフロード処理用コマンドであると判別し、演算処理プログラム3096に処理を渡す。演算処理プログラム3096では、リストとして渡されたIDに対応するディスクごとに、ベリファイを行うためのコマンドをモジュール内データバッファ301に用意する。ベリファイにはディスクのセクタごとに設けられたECC情報のみをチェックする方法と、セクタ全体を読み出せるかチェックする方法とがあるが、前者のチェックはセクタに格納されたデータを読み出せるかどうかをチェックの対象とはしないため信頼性は低い。

#### 【0061】

ここでの説明では、後者の方法によるセクタ全体に対するベリファイを考えるが、セクタ全体をチェックするベリファイコマンドはATAコマンドに存在しない。そこで、セクタ単位でのリードを試み、かつ失敗した際には再試行を試みずにすぐにエラーを返すREAD SECTORS (NORETRY) コマンドによりデータがセクタごとに読み出せるかチェックすることにより同等の機能を実現する。

#### 【0062】

このコマンドが全ての対象ディスクに対して実行されるように、コマンド実行状況管理テーブル3092に登録が終わると、このコマンドリストをモジュール内データバッファ301に用意し、ATAインターフェースコントローラ306に指示を出してコマンドを発行する。コマンドに対して全てのディスクから応答が返り、応答の中にエラーがなければ、上述した方法と同様の手順によって、各ディスクの次のセクタを検査する。これをディスクの領域の最後まで繰り返す。エラーが発生した場合には、そのディスクとセクタ位置を特定できるエラーレポートを作成し処理を続行する。

#### 【0063】

全ての領域の検査が終われば、演算処理プログラム3096はディスクアレイコントローラ102xに報告して、オフロード処理を完了する。

#### 【0064】

以上説明したように、本発明の第1の実施形態に係る磁気ディスクアレイ装置の概要を図11を参照して再度説明する。図11は本発明の実施形態に係る磁気ディスクアレイ装置におけるディスクアレイコントローラ及び処理オフロード機能モジュールの処理態様の概要を説明する図である。

#### 【0065】

ディスクアレイコントローラは、ホストからのコマンドとデータに基づいて、通常（標準）のコマンド（例えば、リード、ライト）と、ベンダユニーク（製造元独自）なコマンド（例えば、パリティ演算付きライト、RAIDフォーマット、オンライン複数ベリファイ、ディスク間コピー（図4参照））とを生成して、処理オフロード機能モジュールに出力する。処理オフロード機能モジュールは、図5に示すように、コマンド解析処理プログラム、演算処理プログラム、及びマッピングテーブルなどを備えており、ディスクアレイコントローラからのFCコマンドをコマンド解析処理プログラムによって解析して、ATAコマンドにマッピング可能か否かを判断し、標準FCコマンドの場合には、マッピング



テーブルから対応するATAコマンドを選択してATAディスクに当該ATAコマンドを発行する。

#### 【0066】

図11において、ベンダユニークな4つのコマンドの中で、RAIDフォーマットとオンライン複数ベリファイのコマンドは、ディスクアレイコントローラベンダユニークなオフロード処理コマンドであり、パリティ演算付きライトとディスク間コピーのコマンドは、ディスクベンダユニークなオフロード処理コマンドである。なお、ディスクアレイコントローラからのディスクベンダユニークなFCコマンド（ディスクアレイコントローラからのFCコマンドはFCディスクのIDを指定している）は、ディスク群にFCディスクがあれば、機能モジュールで処理されることなくFCディスクにコマンド発行されることとなる。

#### 【0067】

ベンダユニークなFCコマンドはコマンド解析処理プログラムによって1対1のATAコマンドにマッピングできないと判断されて、このベンダユニークFCコマンドは機能モジュール内の演算処理プログラムに引き継がれて、図6に示すフローにしたがって演算処理される。図11におけるパリティ演算付きライトコマンドの例で説明すると、旧データと旧パリティのREAD (RETRY) がATAディスクに発行され、次に、パリティ演算されて新データと新パリティがATAディスクにWRITE (RETRY) される。図11において実線はデータの流れを示し、破線はステータスの流れを示す。

#### 【0068】

本実施形態において、ATAディスクにはIDが付加されていないので、サブコントローラである処理オフロード機能モジュールがATAディスクを管理することになる。ディスクアレイコントローラは配下のIDに対してコマンドを発行するため、ディスクコントローラからのコマンドは一旦サブコントローラが全て受け取って処理した後にATAディスクにコマンドを発行することになる。換言すると、ディスクコントローラは配下のディスクIDに向かってコマンドを発行しているのであるから、サブコントローラの存在は見えていない（換言すると、サブコントローラはディスクアレイコントローラに対して透過的にディスクを見せる機能を奏する）ことになる。

#### 【0069】

一方、従来技術においては、ディスクアレイコントローラ配下にサブコントローラとFCディスクがFCループを形成していて、このFCループ上にサブコントローラとFCディスクとがそれぞれIDを有しているので（FC規格上で1ループ当たり127台のサブコントローラとFCディスクが設置可能）、ディスクアレイコントローラはIDで管理しているので、サブコントローラとFCディスクを別々に意識してこれらにそれぞれコマンド発行する必要がある、場合によってはコマンド実行順序が保証されない（例えば、FCディスクからリードした後にライトすべきところを、先にFCディスクにライトするという不都合が生じ得る）。また、従来技術では、サブコントローラとFC-ATA変換部とが別個に縦続接続されていたので、サブコントローラでオフロード処理のFCコマンドが発せられても、FC-ATA変換部において、入力されたFCコマンドに1対1のATAコマンドが対応しない場合が多々あった。

#### 【0070】

本発明の実施形態においては、処理オフロード機能モジュールがその内部に、図3に示すように、コマンド解析処理プログラム、演算処理プログラム、ATAコマンドマッピングテーブル等を備えて、図5と図6の処理フローを実行することで、オフロード処理を内部でATAコマンドに展開することができて、ATAコマンドには直接マッピング不可能なFCコマンドが変換利用できることとなる。また、本実施形態の機能モジュール（サブコントローラ）がFCループIDとATAディスクの対応付けを管理するため、機能モジュールが自分用のIDを持つ必要が無く、即ち、ディスクアレイコントローラが管理の対象とするIDを、機能モジュールで使用（消費）することがない。また、本実施形態では、ディスクアレイコントローラは、標準コマンドもディスクベンダユニークコマンド（パ

リティ演算付けライト、ディスク間コピー)もコントローラベンダユニークコマンド(RAIDフォーマット、オンライン複数ベリファイ)も、サブコントローラとディスクとを意識し分けしてそれらのコマンドを発行する必要がなく、対象ディスクのIDに対して発行すれば済む(機能モジュールが全てのコマンドを受け取ってディスク管理するので)。

#### 【0071】

このように、本発明の第1の実施形態に係る磁気ディスクアレイ装置によれば、オフロード処理用FCコマンドを直接ATAコマンドに変換して処理するため、コマンド変換によるオーバーヘッドを最小にして実行することができる。また、オフロード処理の内容に応じて最適なATAコマンドを用いて処理をすることが可能である。また、本実施形態と同様の方法によって、ドライブ間のデータコピーや複数ディスクにRAIDフォーマット処理も行うことが可能である。

#### 【0072】

##### 「第2の実施形態」

図7は、本発明の第2の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールを磁気ディスク格納筐体のキャニスタに実装した構成を示す図である。図7において、80xはキャニスタ(ディスク収納容器)を表し、810xは各キャニスタ内に組み込まれた処理オフロード機能モジュールを表す。

#### 【0073】

図8は処理オフロード機能モジュールをキャニスタ内に実装した場合の内部構成を示す図である。8200は第2の実施形態に関する処理オフロード機能モジュール810xを搭載したコネクタボードであり、磁気ディスク格納筐体側接続インターフェース8201とディスク側接続インターフェース接続用インターフェース8202を備える。

#### 【0074】

本発明の第2の実施形態によれば、上述した第1の実施形態と同様の効果を得られる上に、ディスクごとにオフロード処理が可能となる。オフロード処理で生じるデータの移動はキャニスタ内に限定されるため、データのバス効率はよりよくなる。また、処理の並列多重度が上がり、ディスクアレイコントローラやアレイコントローラにつながるバス上にほとんど負荷をかけることなく単位時間当たり処理性能を向上させることが可能である。

#### 【0075】

##### 「第3の実施形態」

図9は、本発明の第3の実施形態に係る磁気ディスクアレイ装置における処理オフロード機能モジュールのブロック図であり、FCディスクとATAディスクを混載可能とした場合の構成を示す図である。図9において、同一の磁気ディスク格納筐体104x内に、収めるATAディスクの一部をFCディスクに置き換えて混在させた場合における処理オフロード機能モジュール106の内部構成を示している。

#### 【0076】

本発明の第1の実施形態の構成に比べて、混在を実現するために、機能モジュール106はFCディスク970xを接続するためのFC接続インターフェース908xと、このインターフェースをコントロールするためのFCインターフェースコントローラ906と、オフロード処理を適切なFCコマンド群に展開するためのFCコマンドマッピングテーブル9098と、を新たに備えている。

#### 【0077】

本発明の第1の実施形態との動作上の差異について以下説明する。FCインターフェースコントローラ303が上位のFC接続インターフェース302xを通じてコマンドを受け取ると、コマンド解析処理プログラム3093はディスクID情報管理テーブル3091を参照して、そのコマンドがATAディスクに対するものなのかFCディスクに対するものなのかを判別する。送られてきたコマンドが通常のFCコマンドであり、かつFCディスクを対象としているのであれば、当該FCコマンドを変換することなく素通りさせ、

A T A ディスクが対象であった場合は、A T A コマンドマッピングテーブル 3096 を用いて、最適な A T A コマンド群に変換する。

【0078】

また、受け取ったコマンドがオフロード処理用コマンドだった場合、コマンド解析処理プログラム 3093 は演算処理プログラム 3097 に処理を移し、対象ディスクの種別と処理内容に応じて最適なコマンド群に変換して処理を行う。

【0079】

本発明の第3の実施形態によれば、第1の実施形態で得られる効果に加え、異なるディスクを同一の上位コマンドで操作することが可能であり、また、その上位コマンドをディスクの種類と処理内容に応じた最適なコマンド群を用いて処理することが可能である。

【0080】

「第4の実施形態」

図10は、本発明の第4の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールをディスクアレイコントローラに実装した構成を示す図である。図10にはディスクアレイコントローラ 102x に処理オフロード機能モジュール 1103x を実装した場合の磁気ディスクアレイ装置の構成を示している。

【0081】

処理オフロード機能モジュール 1103x を搭載するディスクアレイコントローラ 102x は、複数の A T A ディスクをスイッチ的に接続する A T A 接続部 11110 と、同様の方法によって F C ディスクを接続する F C 接続部 11120 とを介して、A T A ディスク 17xx と、F C ディスク 110xx とに接続される。なお、図10において、x、y、z、a、b、m、n は任意の自然数である。

【0082】

また、図10では、A T A ディスクと F C ディスクを挙げているが、これ以外のディスクの組み合わせでもよい。例としては、A T A ディスクと、S A S (S e r i a l A t t a c h e d S C S I) ディスクの組み合わせ等であってもよい。また、磁気ディスクの種類は2種類以上であってもよい。

【0083】

本発明の第4の実施形態によれば、ディスクアレイコントローラは一種類のコマンドが発行できればよく、下位に接続されるディスクのインターフェースごとにコマンドを用意する必要がなくなり、上述した第3の実施形態と同様に、ディスクに応じた最適なコマンド処理を実行することが可能である。また、処理オフロード機能モジュールはディスクアレイコントローラ上に実装されるため、ディスクアレイコントローラ処理部(図2に示す構成と同様な構成)と処理オフロード機能モジュール間のプロトコルは、ベンダユニークな第3のコマンド体系であってもよい。

【図面の簡単な説明】

【0084】

【図1】本発明の第1の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールを磁気ディスク格納筐体のインターフェースカードに実装した構成を示す図である。

【図2】本発明の第1実施形態におけるディスクアレイコントローラの内部構成を示す図である。

【図3】本発明の第1の実施形態に関する処理オフロード機能モジュールの内部構成を示す図である。

【図4】図2に示すディスク情報管理テーブル 2063 の詳細構造を示す図である。

【図5】本発明の第1の実施形態に係る磁気ディスクアレイ装置におけるディスク情報取得処理の動作を示すフローチャートである。

【図6】本発明の第1の実施形態に係る磁気ディスクアレイ装置において、複数のディスクに異なるコマンドを送るオフロード処理例としてのパリティ演算付きライトの

動作を示すフローチャートである。

【図 7】本発明の第 2 の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールを磁気ディスク格納筐体のキャニスタに実装した構成を示す図である。

【図 8】本発明の第 2 の実施形態に関する処理オフロード機能モジュールをキャニスタ内に実装した場合の内部構成を示す図である。

【図 9】本発明の第 3 の実施形態に係る磁気ディスクアレイ装置における処理オフロード機能モジュールのブロック図であり、F C ディスクと A T A ディスクを混載可能とした場合の構成を示す図である。

【図 1 0】本発明の第 4 の実施形態に係る磁気ディスクアレイ装置における全体構成のブロック図であり、処理オフロード機能モジュールをディスクアレイコントローラに実装した構成を示す図である。

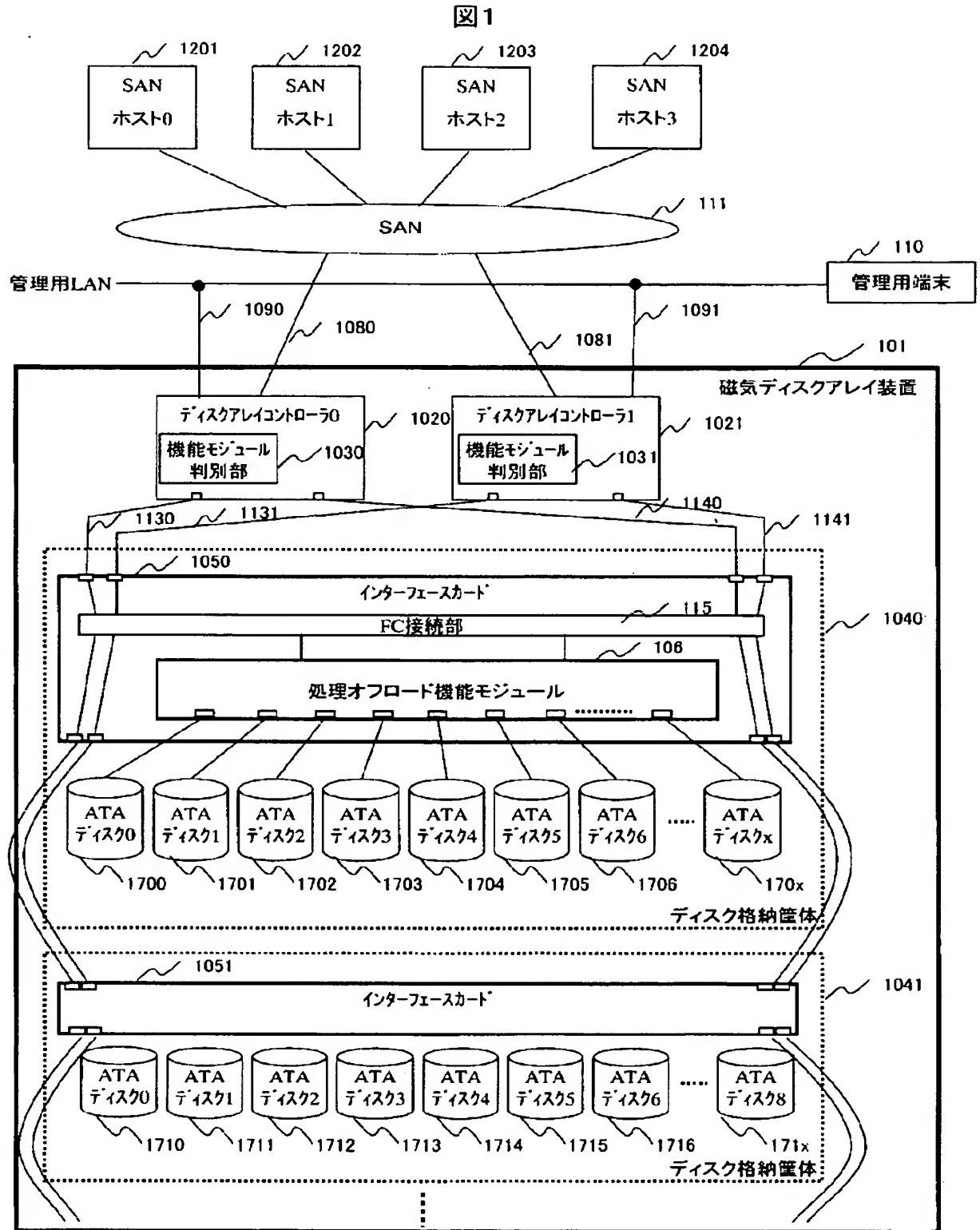
【図 1 1】本発明の実施形態に係る磁気ディスクアレイ装置におけるディスクアレイコントローラ及び処理オフロード機能モジュールの処理態様の概要を説明する図である。

【符号の説明】

【 0 0 8 5 】

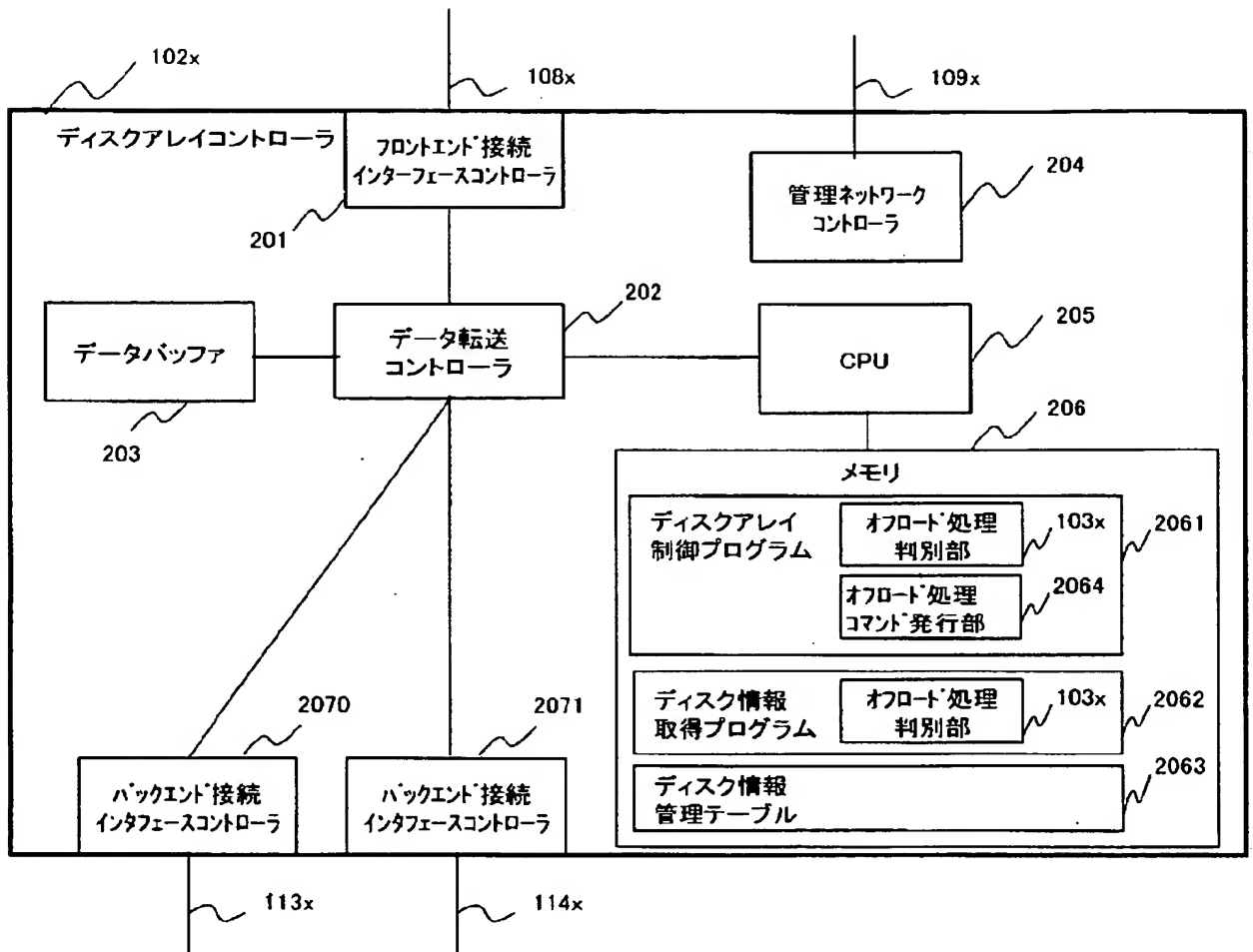
- 1 0 1 磁気ディスクアレイ装置
- 1 1 0 管理用端末
- 1 1 1 S A N (ストレージエリアネットワーク)
- 1 1 5 F C 接続部
- 1 0 6 処理オフロード機能モジュール
- 1 2 0 x S A N ホスト
- 1 0 2 x ディスクアレイコントローラ
- 1 0 4 x 磁気ディスク格納筐体
- 1 0 5 x インターフェースカード
- 1 7 0 x, 1 7 1 x A T A 磁気ディスク
- 2 0 6 1 ディスクアレイ制御プログラム
- 2 0 6 2 ディスク情報取得プログラム
- 2 0 6 3 ディスク情報管理テーブル
- 3 0 9 2 コマンド実行状況管理テーブル
- 3 0 9 3 コマンド解析処理プログラム
- 3 0 9 4 ステータス処理プログラム
- 3 0 9 5 A T A コマンドマッピングテーブル
- 3 0 9 6 演算処理プログラム
- 8 0 x キャニスタ
- 8 2 0 0 コネクタボード
- 3 1 0 x A T A 接続ライン
- 9 1 0 x F C 接続ライン

【書類名】 図面  
【図 1】



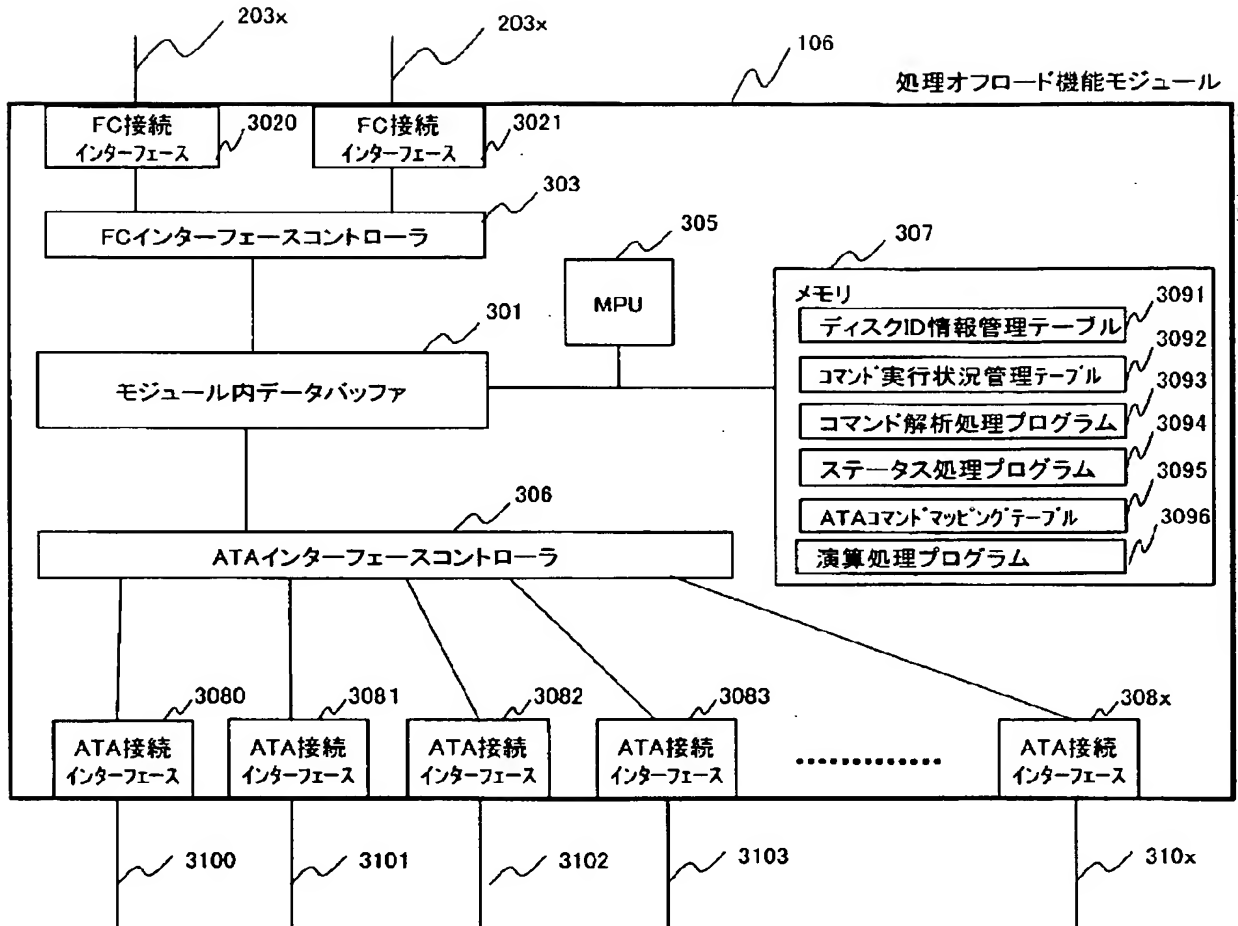
【図 2】

図2



【図 3】

図 3



【図 4】

図 4

2063

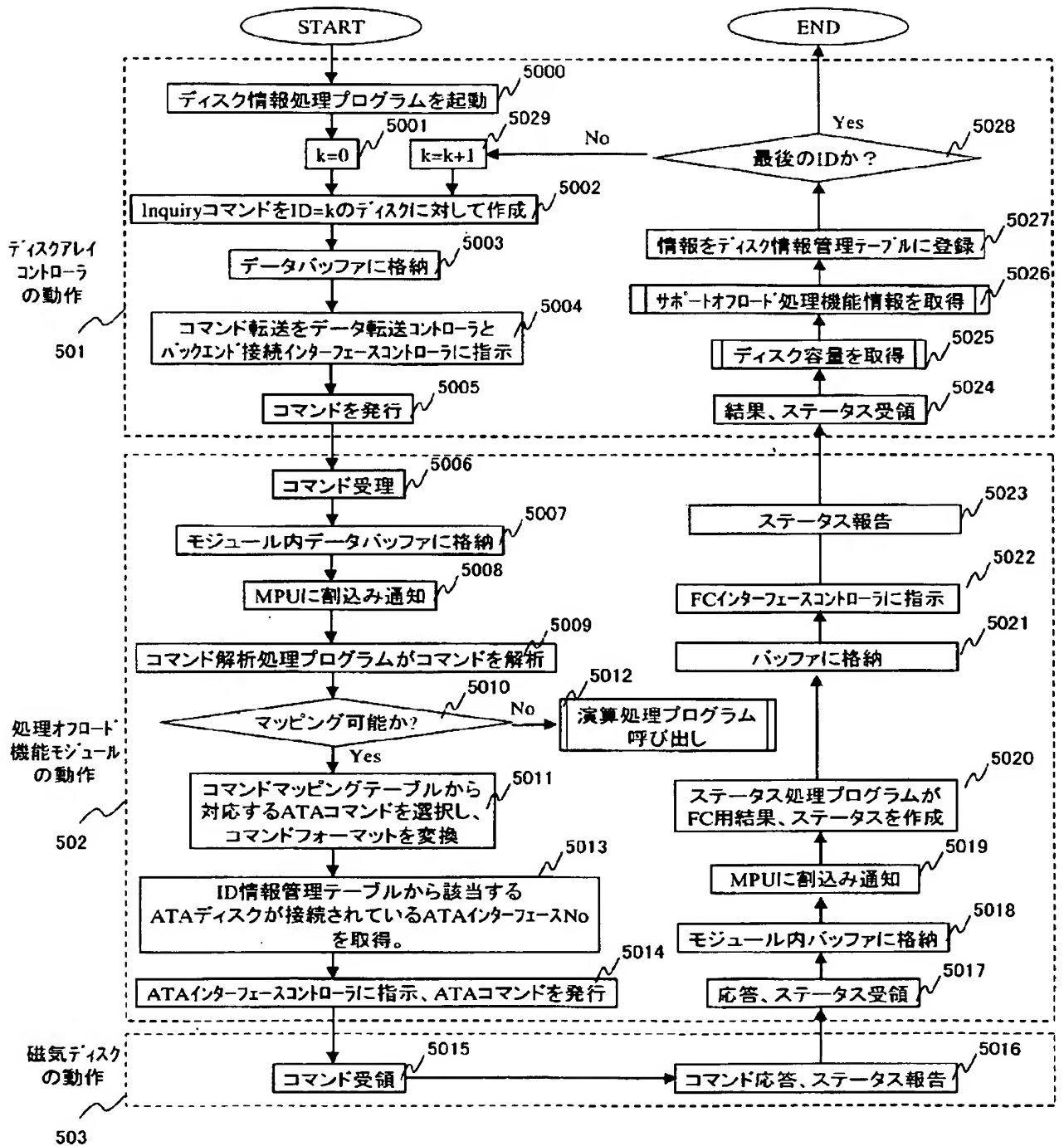
ディスク情報管理テーブル

No	ベンダ 名	モデル名	シリアルNo	容量	状態	オフロード機能			
						パリティ演 算付ライト	RAID フォーマット	オンライン ヘルファイ	ディスク コピー
#0	Hitachi	xxx-xxxxxxx	7F0450370A	200GB	良好	○	○	○	○
#1	Hitachi	xxx-xxxxxxx	7F0538432A	200GB	良好	○	○	○	○
#2	Hitachi	xxx-xxxxxxx	7F327532EB	200GB	良好	○	○	○	○
#3	Hitachi	xxx-xxxxxxx	8F3489601A	200GB	良好	○	○	○	○
#4	Hitachi	xxx-xxxxyyy	3W6014601D	200GB	良好	○	○	○	○
#5	Hitachi	xxx-xxxxxxx	2A3260112E	200GB	良好	○	○	○	○
#6	Hitachi	xxx-xxxxxxx	8E5436292F	200GB	良好	○	○	○	○
#7	Hitachi	xxx xxxxxxx	3A5236017F	200GB	良好	○	○	○	○
#27	Hitachi	xxx-xxxxyyy	3H2160143D	180GB	良好	○	×	○	○
#28	Hitachi	xxx-xxxxyyy	3B3541128A	180GB	良好	○	×	○	○

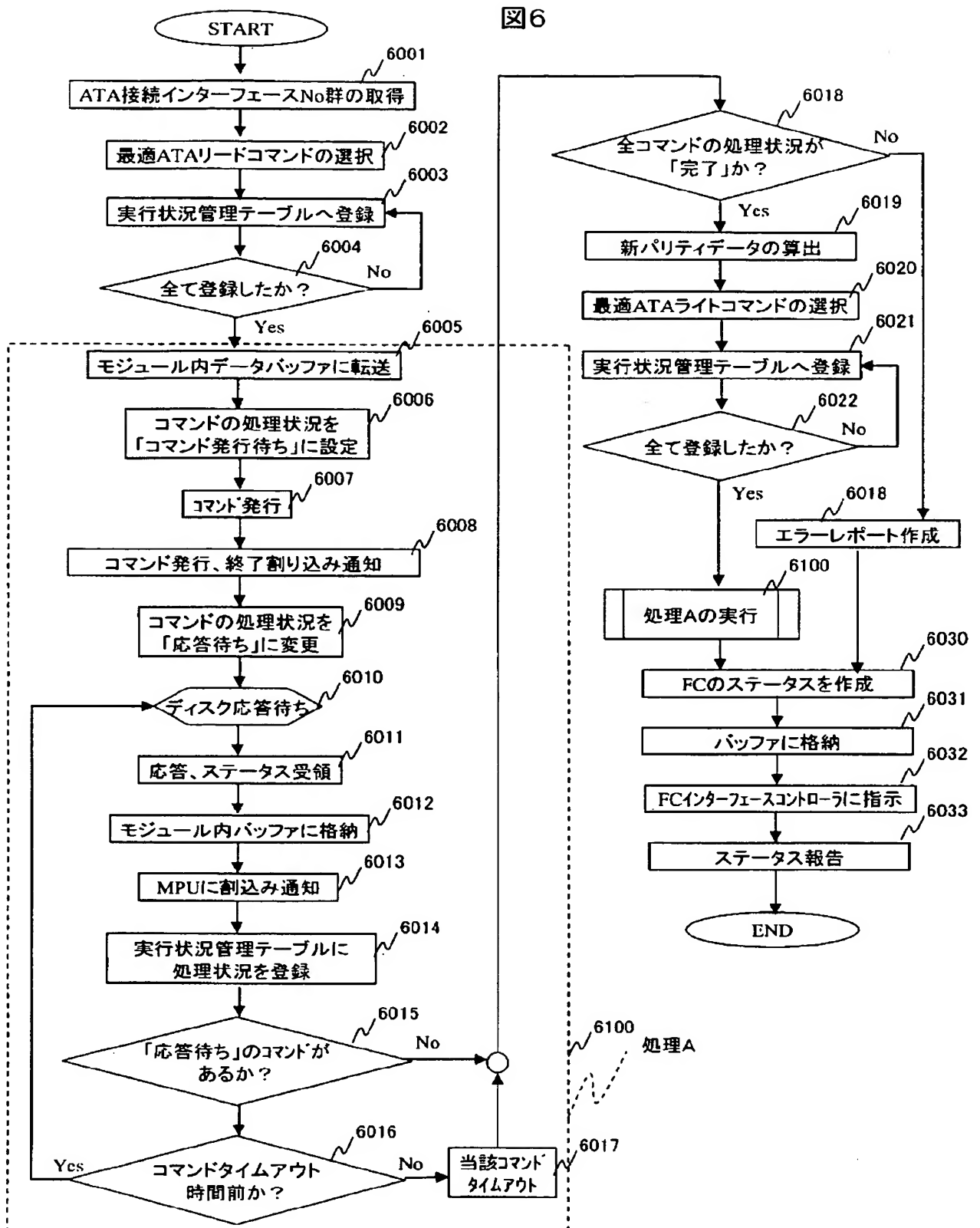


【図 5】

図 5

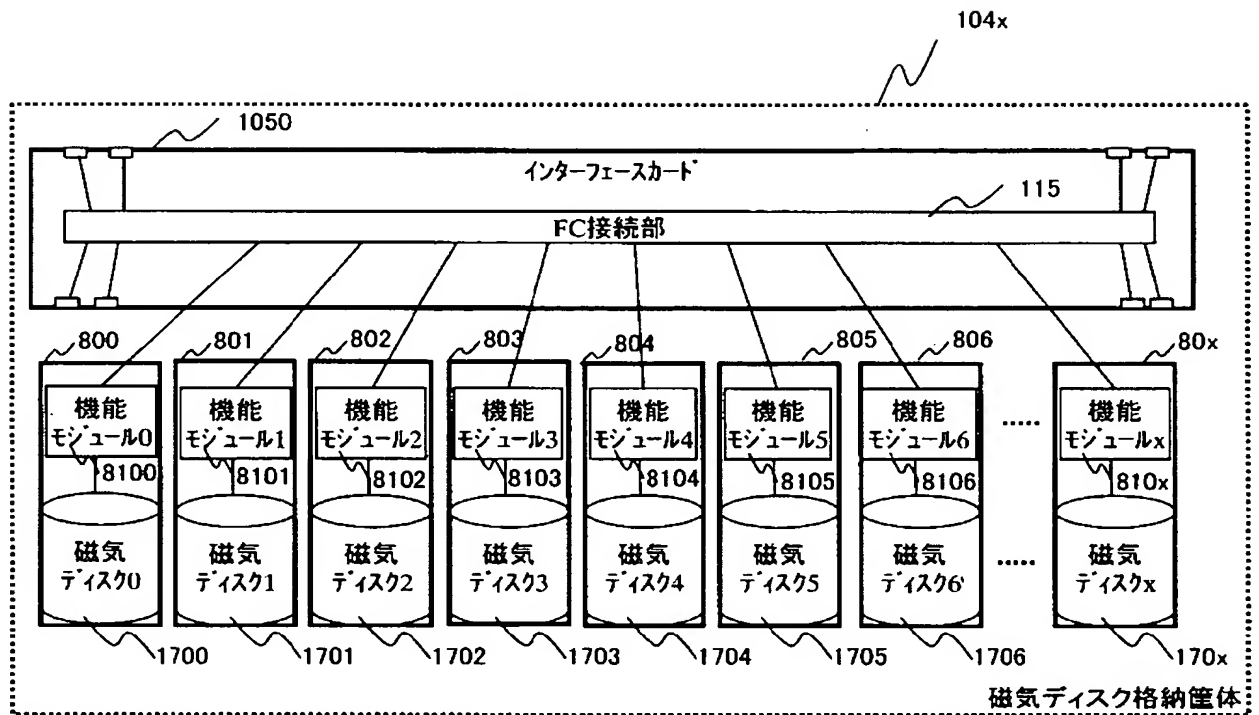


【図 6】



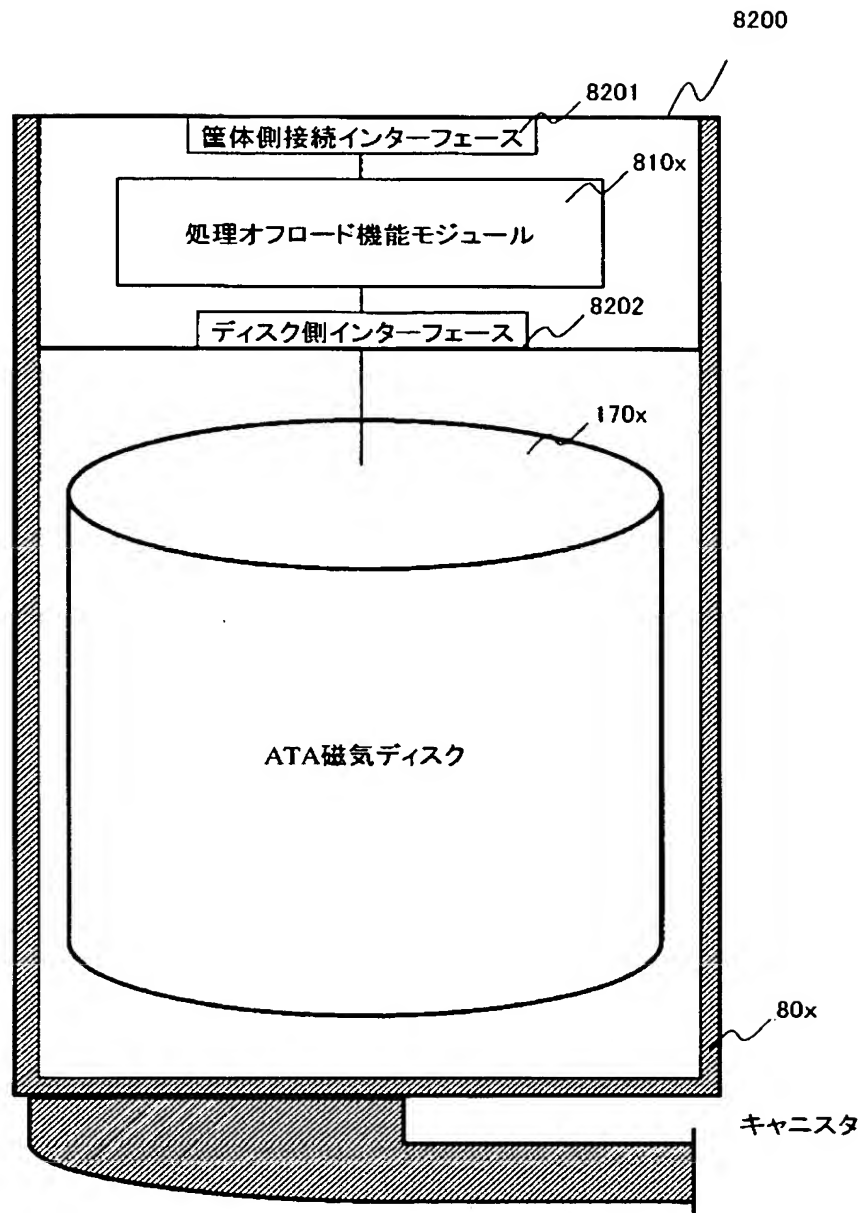
【図 7】

図 7



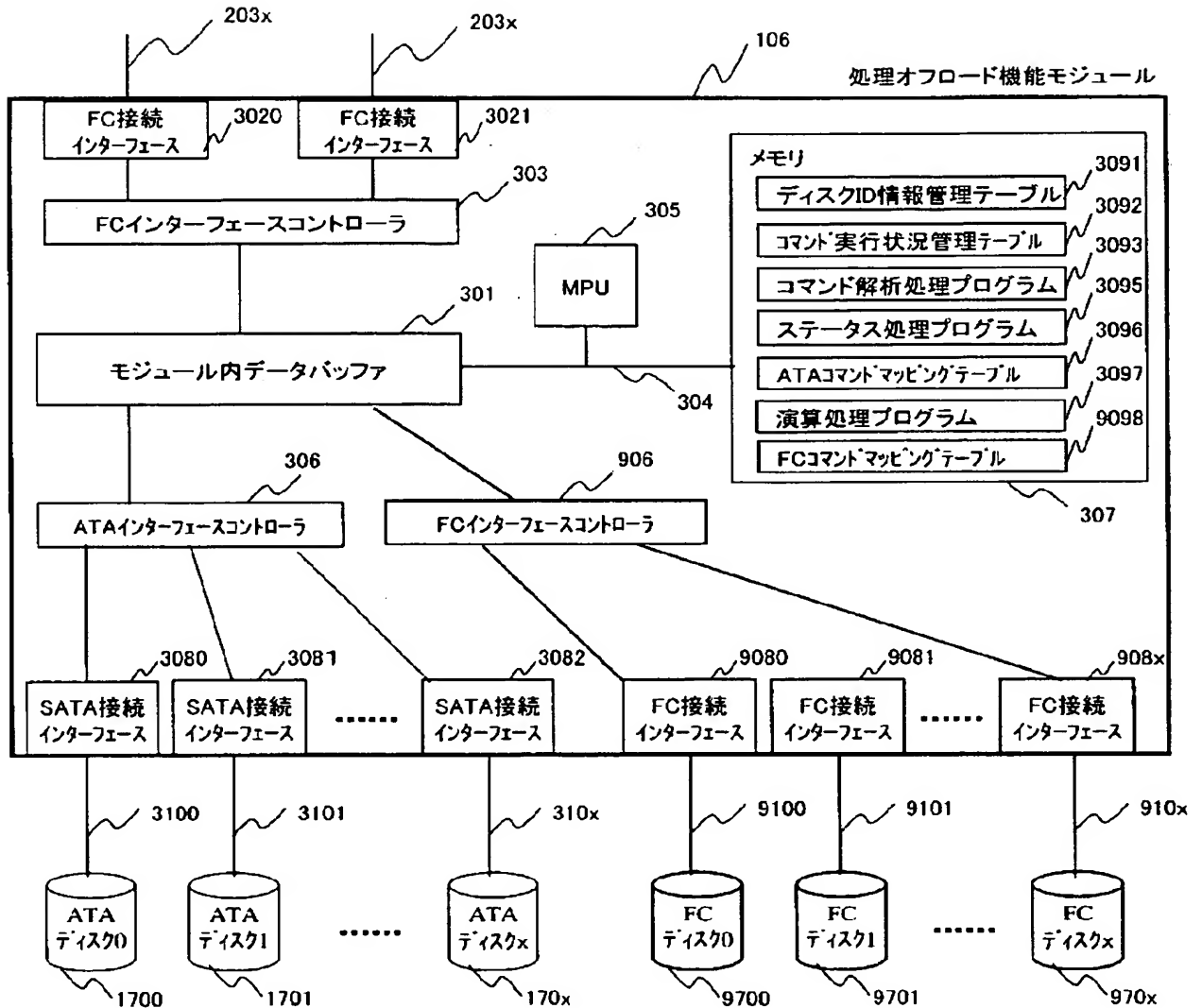
【図 8】

図 8



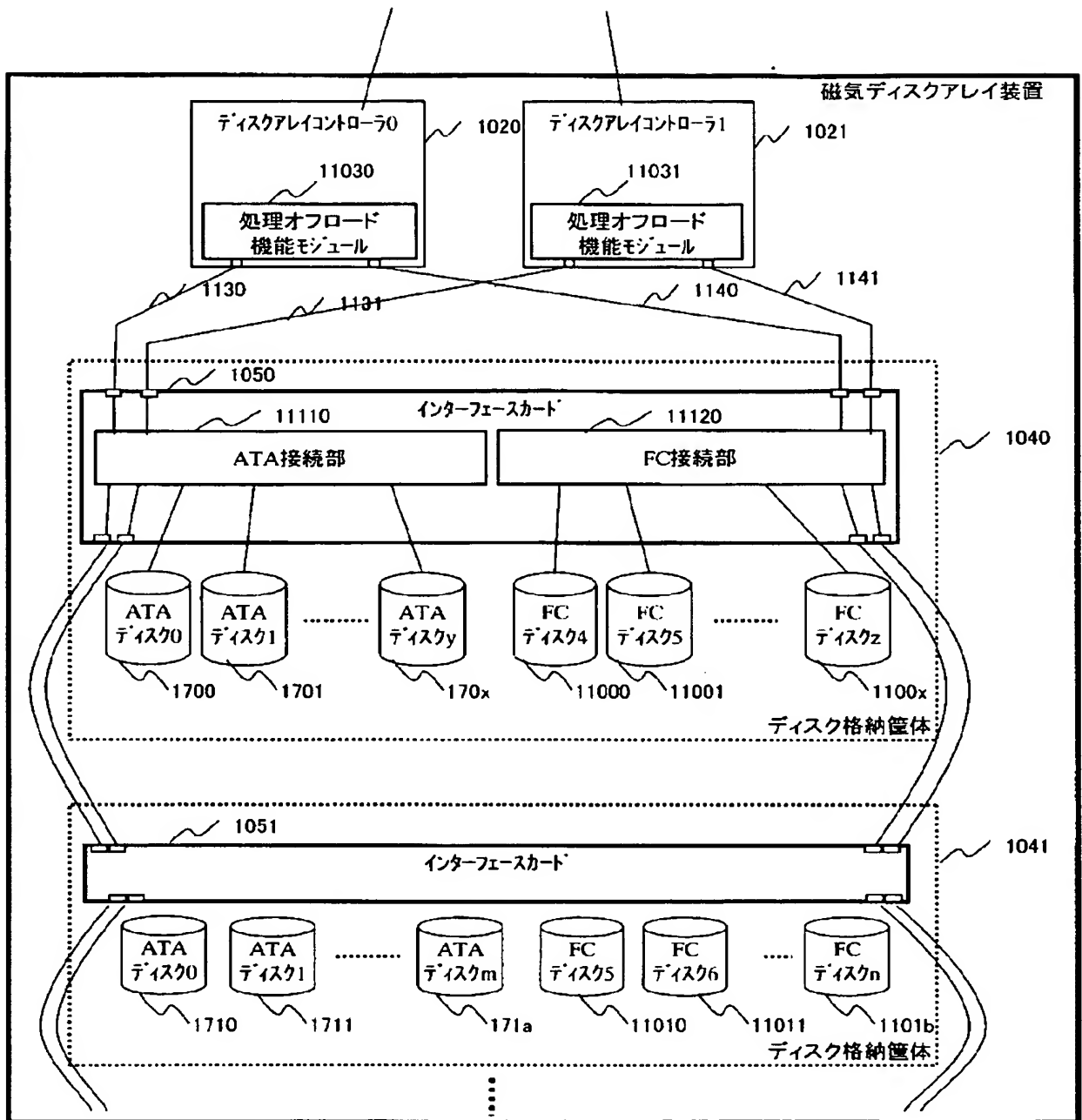
【図 9】

図 9



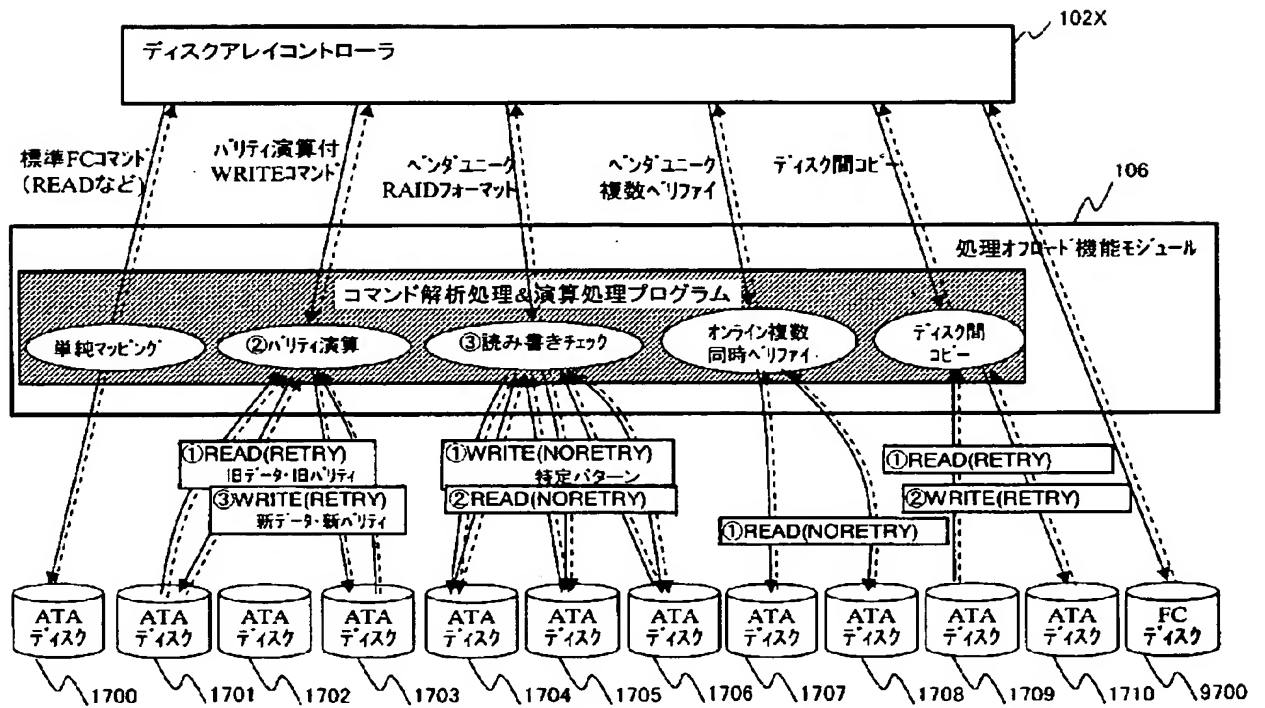
【図10】

図10



【図 11】

図 11



**【書類名】 要約書****【要約】**

**【課題】** A T A ディスクの利用を可能としつつ、F C - A T A コマンド変換によるオーバーヘッドを削減し、最適な A T A コマンドを生成し、I D 等のリソースを無消費にして、オフロード処理を実現すること。

**【解決手段】** A T A 磁気ディスクと、A T A 磁気ディスクを制御するディスクアレイコントローラと、ディスクアレイコントローラと A T A 磁気ディスクとの間の処理オフロード機能モジュールと、を備えた磁気ディスクアレイ装置であって、ディスクコントローラは、リード又はライト等の処理を行う標準処理 F C コマンドと、ベンダユニークなオフロード処理を行うオフロード処理 F C コマンドと、を出力し、処理オフロード機能モジュールは、標準処理 F C コマンドに対しコマンドマッピングして対応する A T A コマンドを A T A 磁気ディスクに発行し、オフロード処理 F C コマンドに対し A T A プロトコルで最適処理となる A T A コマンド群を用意する。

**【選択図】 図 1**



特願 2 0 0 3 - 3 9 3 9 1 2

出 願 人 履 歴 情 報

識別番号

[ 0 0 0 0 0 5 1 0 8 ]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所